Use of OAI protocol and its impact in digital libraries: a case study in Spain, Portugal and Latin America[1]

Fernanda Peset [1], Antonia Ferrer [1], Imma Subirats[2] and Adolfo Muñoz[1]

[1] Polytechnic University of Valencia, Spain
{mpesetm, anfersa, amunyoz}@upv.es

[2] Food & Agriculture Organization (FAO), Rome, Italy
imma.subirats@fao.org

Abstract. The current communication approaches the situation of the development of repositories that use the Open Archives Initiative (OAI) protocol for data collection. These type of digital libraries are undergoing a worldwide boom. This work studies the current state of their implantation in Spain, Portugal and Latin America. In order to do so, the existing projects in the official records have been studied using a methodology beyond the existing records of repositories. It concludes that the situation in 2006 is fairly encouraging in so far as the number of projects, but that it is quite deficient as concerns the quantity of data stored

Keywords: Digital Libraries, Open Archives Initiative, Open Access, Spain, Latin America, Portugal, Scientific communication, Indicators

1 Introduction

The information sector is currently undergoing remarkable changes determining the appearance of new methods and elements for managing digital contents. Moreover, with the introduction of technology and of the networks of communication, an incredible number of facilities for distribution and reproduction appear. As a result, other concerns for the management of the information arise, such as: the respect for the rights of the producers, the protection of the authors' rights, the new means of circulation of contents, or the need to preserve the digital heaps. As such, the definition of digital library may currently prove to be widespread and extensive. It is hard to concretize what type of objects contained in the compilations, that is to say the formats

[1] The presentation of this communication in the Conference has been funded by the Programa de Apoyo a la Investigación y Desarrollo (PAID-04-06), Universidad Politécnica de Valencia.

ARD Prasad & Devika P. Madalli (Eds.): ICSD-2007, pp. 459-471, 2007

that sustain the information, the origin of the metadata, or who conserves the digital object.

Added to the need to make what already exists accessible and to exploit the contents, is that of adding value to the services lent with information (Peset & Prats, 2006). If the compilations are already digital, it will be easier to create new services. The use of inter-operable technologies allows the re-use of contents for other services avoiding routine tasks such as the assignation of metadata. This re-use is one of the bases of the Open Archives Initiative given that the archives or providers of services store and conserve the digital objects whilst the providers of services collect this data in order to create services with added value. For example, DARE offers a service which presents the results of the most important Dutch scientists i.e. The Cream of Science, at http://www.creamofscience.org. Using the power of this protocol, any service with added value could be designed: services for suggested reading material (such as Amazon), indexes of references (CiteBase), other users' comments, or a service that offers the data in different forms of bibliographical reference. The need to preserve the digital assets exists, both those that have been digitalized and those that are born digital. The quality information must be conserved, be it for its hereditary character or for its scientific usefulness. The speed of the technological change shows the necessity not to depend on one concrete situation. A good example can be found in the norm *Reference Model for an Open Archival Information System* (OAIS) ISO 14721:2003 which has established itself as the reference model for the management of information systems, for transferring data and also for conserving it. It defines models and common terminologies for dealing with the changes.

Definitively, the present work aims to approach the outlined issues on the basis of the solutions presented by the OAI for the distribution, management, preservation, certification etc. of the scientific literature (Fernández, 2005). The open access to scientific information is a hot potato in the university environment and research centres (Barrueco & Subirats, 2003). Its benefits are based on the freedom of choice of software, the development towards new models for scientific communication, and the use of data stored in different repositories with the aim of offering information services with an increased value. They are, moreover, visible to the Internet browsers and form a part of the semantic web thanks to the introduction of descriptive metadata. Due to these benefits, the OAI archives prove to be a forerunner for the introduction of digital libraries (Ferrer, Peset, Moreno & Lloret, 2005).

In the context of Open Access, two models or paths have been introduced in order to reach the objectives: either to publish in the journals of open access combined in the *Directory of Open Access Journals* (DOAJ), or to publish in whatever journal but store a copy in open access repositories freely accessible

on Internet. The creation of stores increases the visibility, dissemination, access, and in some cases the reference to the research (Subirats, 2004 and Harnad & Brody, 2004). In another way, it forms a source of information which is fundamental to all of those centres, particularly in developing countries, that lack the necessary resources to have sources of information in full text format in their libraries. In recent years, the growth in the number of documents stored has been exponential to their registration in international records. Nevertheless, the diversity in the contents of every one of the initiatives recorded leads to a confusing panorama as regards their impact as digital libraries. The proliferation of such stores is making clear the necessity to improve the tools for identification and localization of the same as well as another kind of actions to promote the implantation of Open Access.

2 OAI and its impact in digital libraries in Spain, Portugal and Latin America

Spain, Portugal and Latin America are united by historical, cultural and linguistic ties which makes them suitable as an object for a single study. There are relationships between the professional communities which go beyond the simple geographical proximity – peninsular and continental respectively. This is the reason why we have made this the object of our study. This work aims to lay the foundations of the impact that OAI is having in Spain, Portugal and Latin America through a systematic and exhaustive study of the actions that are being carried out (Serrano & Prats, 2005 and Frías & Santiago, 2006). The fact that there is no record of said projects makes it extremely difficult to localize them and to evaluate the reach of the open access in these geographical areas. The lack of basic data makes it impossible to share experiences between the projects and limits their visibility. This work is, definitively, the prior study for the development of the OA-LATINO project, the reference web for Hispano-Luso speaking OAI. The results will be the scientific base for the laying down of recommendations for good practices to promote the real access to the whole texts of the publications in the digital libraries.

We begin with the certainty that these geographic areas are in an incipient moment, for which reason strategies may be designed which will allow the maximum advantage to be obtained from the efforts invested in the sector of e-science. The application of the protocol is fairly well-known, but depending on the sector in which we find ourselves (documentation, computing, university management...), it has advanced more or less. There is proof of the existence of a number of repositories, but this does not mean that the application of the OAI protocol is providing a wider visibility to the scientific works. According to Oliver & Swain (2005) the aim of monitoring "…for

growth in numbers of repositories, their compilations and content development, the services they provide, their acceptance and use by scholars, and their impact on scholarship" has been only partially possible. The data collected does not allow the measuring of all those aspects that would be desirable.

2.1 Methodology

Since May 2006, a compilation has been made of all of the repositories and providers of services created and maintained in Spain, Portugal and Latin American countries. The main sources of information existing on the web were identified: The OA records (ROAR, OpenArchives & OpenDOAR). They were completed using the data of Oliver & Swain (2006).

Secondly, the lists of data providers compiled by the providers of services or compilers were used. To discover new service providers the records of Open Archives and the University of Illinois (UIUC) were used. All of the information they held of Spanish or Brazilian and of international ambit were noted: *OAIster, SDL: Search Digital Libraries, Public Knowledge Project, Open Archives Harvester, ROAI, REBIUN, E-Ciencia* and *HOLMES-Descobrindo a Ciência da Informação por você!*

The results of the analysis were stored in a database called refOA. For each and every one of the elements localized, the following fields were described: *Name, Nature, Country, URL, Nº of 12/2006 Documents, Year started, Type of content, Promoting institution, Software, Contact name, Record, Nº personnel, Procedure Manuals, Policies for entry of information, Descriptive metadata, Classification system/Indexing, Data sets,* and *Region.* After compiling the information appearing in every one of the resources identified, contact was made with the person in charge of the projects in order to confirm the data collected. To reach the first milestone, the project got in touch with those projects localized in Spain. Once the database contained all of the information compiled, a brief survey was sent to the two most important Spanish mailing lists in the ambit of Library and Information Science and Open Access respectively: IWETEL & OS-Repositories. The survey helped to localize those projects that were not until now sufficiently visible on the Web and in the OA ambit, as well as to correct the errors arising in the prior compilation of data. Other initiatives were also found through the bibliography, reports such as that of the Spanish Web of University Libraries (Rebiun y Cobacho, 2006), or through personal contacts etc. It should be borne in mind that in the immediate future the second milestone will be reached, applying this methodology to the Portuguese and Latin American initiatives.

2.2 Results

The total number of actions identified is 150. Among these are counted not only repositories and services/compilers, but also all kinds of actions related with the implantation of the OAI & Open Access protocol, or that are at least recorded in the international records mentioned previously. As shall be seen, actions for integration whose aim is to share knowledge among the administrators and professionals linked to the projects are also counted, as would be the case of the portals or wikies -Grupo de Usuarios de Dspace in Spain. As has already been mentioned, this level of exhaustiveness has been possible in the case of Spain thanks to the help of a number of professionals who have seen the project we are describing as an opportunity to give a greater visibility to the work performed in Spain. The works of Rebiun (2006), Frías (2006) or Prats (2006) especially have allowed the existence of numerous initiatives being developed in Spain to be detected.

The total number of providers of data currently in an embryonic state is 20, and as such the data presented will, in general, refer to 130 actions. The following is the distribution according to the nature of the action:

Type	**Number of actions**
Working groups/User groups	2
Mailing lists	2
Portals	1
Data providers (DP)	83
Service providers (SP)	10[2]
Journals	34
Total	**132**[2]

Table 1 Distribution of the number of initiatives by type

The initiatives that do not appear in records, as are two mailing lists or working groups, are Spanish whilst the portal of access in Portuguese is of Brazilian origin. In so far as the geographical distribution, the initiatives situated in Latin America and Portugal amount to 82, whereas there are 43 Spanish ones, and the international initiatives with Latin American or peninsular participation add up to 5. Some type of activity has been identified in twelve countries, all –save Costa Rica & Uruguay- with a Gross Domestic Product (GDP) superior to 85,000 Billion US dollars for 2006. To these effects, 24 countries don't exist, all with a GDP inferior to 36,000 and which in general are small Caribbean and Central American countries.

The earliest project is TDX/TDR, a digital library of theses from all of the Spanish universities which began as a project of the Consorcio de Bibliotecas

[2] Two of them are at the same time Data providers and Service providers

Universitarias Catalanas in 2001. In general it is not easy to know the date when the services first became offered, since that which appears in the records often corresponds to the date of inscription. The number of known starting dates is 53, which is less than 50% of the initiatives recorded. Better results have been obtained for Spain, with the starting dates of 38 actions being known –including two internationally ranging projects with a high Spanish participation- which makes for 72% of the total. The distribution allows us to identify a boom which began in 2004 and which we hope will be consolidated in 2007 if the 20 projects claiming to be under development materialize.

Following this brief presentation of results, let us make a detailed analysis of the distribution of the data providers and journals which are, in the end, the most numerous and most well-known initiatives. In so far as the nature of each of the projects, they differ in many aspects. Some distribute only whole texts, others depend on an institution, yet others are a thematic catwalk or the name of a journal.

Country	**Number of projects**	**Nº Data providers**
Argentina	3	3
Brazil	44	14
Chile	3	3
Colombia	3	3
Costa Rica	3	3
Cuba	1	1
Spain	43	28
International	5	5
Mexico	7	7
Peru	4	4
Portugal	8	8
Uruguay	1	1
Venezuela	5	3
Total	**130**	**83**

Table 2 Geographic distribution of the number of actions and distribution of the number of data providers by country

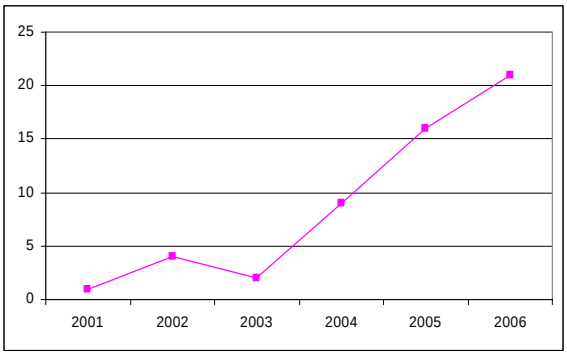

Figure 1. Distribution of the number of known starting dates by year

Spain is the country with the largest number of archives, which we suppose will have repercussions in a smaller average size, although this will be seen later. For their part, the journals served through the OAI-PMH protocol proliferate in Brazil, with 28, whilst in Spain there are 5 and in Venezuela one. The quantity of contents distributed by the data providers and the journals (which after all are data providers) amounts to 354,781 records. It should be pointed out that data have been obtained from 81 of the 117 providers compiled, which is 60%, but not all of them. The distribution of data by country gives the following result:

Country	**Records**	**DP and OA Journals**
Argentina	1.814	3
Brazil	110.632	42
Chile	16.174	3
Colombia	3.967	3
Costa Rica	9.274	3
Cuba	7.176	1
Spain	155.656	33
International	11.625	5
Mexico	4.763	7
Peru	1.744	4
Portugal	16.993	8
Uruguay	7	1
Venezuela	26	4
TOTAL	**339.851**	**117**

Table 3 Distribution of the number of records and providers (archives and journals) by country

It can be seen that 78% of the information is served between Spain and Brazil. The relation between the number of providers and the number of records presents a fairly significant panorama. The average is around 4,000 records per repository. In Cuba only one provider has been identified, concentrating the highest average number of records, whilst the figure for countries like Uruguay or Venezuela is not considered to be representative of the situation.

Below the average are, significantly, Argentina, Mexico or Peru. In the case of the first two, we believe that this has happened because not all the data are available, and in Peru the average size of repositories is small. For the Spanish case, we proposed a smaller average size due to the large number of archives. Nevertheless, the figures do not support this hypothesis. The explanation could be in that in Spain there are two large data providers that do not offer the whole text, but who distribute their metadata (Dialnet).

In order to demonstrate the relation between economic development and implantation of the OAI we shall use the Index of efficiency in the distribution of contents. This index is a new indicator which shows the efficiency of a country by dividing the number of records distributed in relation to the average size of a repository, and relating it to the GDP.

Country	**Index of efficiency**	**Gross domestic product in Billions of US dollars[3]**
Chile	1,9274535	118.188
Cuba	1,710326	35.543
Spain	1,2366337	846.782
Brazil	1,19854557	1.023.198
Costa Rica	0,73678759	20.453
Portugal	0,50626341	1.126.565
Argentina	0,43234829	210.147
Mexico	0,37840407	85.050
Colombia	0,31516459	136.840
Peru	0,13855484	183.619
Venezuela	0,00619683	156.482
Uruguay	0,00166838	17.398

Table 4 Distribution of the index of efficiency in the dissemination of contents by country

This comparative table allows the planning and coordination of the development of the countries in the OA sector. If we accept an index of efficiency of 0.5 as sufficient, the countries with a higher GDP that are below this figure should correct their situation: Argentina, Colombia, Peru and

[3] International Monetary Fund, World Economic Outlook Database, April 2006.

Venezuela. The journals offer far less records (33,688 records, 9%) than the data providers, with a quota of 91%, although they all contain the whole text, which cannot be said of the latter. In any case, these data should be used cautiously since some data providers include journals which are also recorded as such. Perhaps the future of this study, and thus of the OAI initiative, will be to confront the overlapping in the provision of contents. For example, the contents of the international initiatives of Scielo (Scielo.org or the version of science and social sciences) are certainly fed by the thirteen national versions of the Scielo initiative and by the journals that appear in such abundance in Brazil. In this way, the overlapping would modify the data of this study.

Finally we will refer to the service providers, of which we have identified ten initiatives; eight Spanish, one Venezuelan and one Brazilian. The initiatives include various kinds of limitations: geographic, institutional and thematic, for information sciences. These services are made up of harvesters that allow basically the federative searches. There is little more. However their reach and their nature vary widely amongst themselves. It is clear that a service that compiles few archives, or on a concrete theme, has the same use and impact as serving numerous data coming from a number of projects –the case of ROAI or the Rebiun project-. Moreover, two of them are surprising as they harvest metadata that do not appear to come from OAI archives. On the one hand, the Venezuelan case Repositorium serves the data from libraries, we assume all of reference. On the other hand, E-Science feeds basically on UNED videos which are not served as an independent archive.

Finally, what a more detailed analysis of the services has permitted has been to prove the duplication of efforts and the overlapping in the provision of contents. These providers are serving around 400,000 records with there being no evidence of redundancy. To illustrate this overlapping, we have studied the number of coincident or unique archives between the ROAI and Rebiun services, with the result of 70% overlapping between the two.

3 Conclusion

With this study the first milestone has been reached in such a way as that we can now put forward some conclusions. In first place, deficiencies have been detected in the records of the OA ambit. As these records only compile data and service providers, it is difficult to create gateways which do not solely identify the initiatives, which is what these records do. The objective of these gateways would be to share resources, localizing the projects that might interest a specific community. In this way, the generation of tailored-services aimed at the end-user would mean a real step forward respect to the current situation. The fact that the inscription in said records is voluntary, or that it is assigned to only one country means that many of the projects that are

currently emerging become weakened through not being efficiently described. This is the case of the data providers in which a number of countries take part, such as E-LIS or HAL/REDIAL (Peset, Subirats, Barrueco & Noverges, 2003). As an example, we would point out the number of initiatives that ROAR identified as Spanish: just one project in the last analysis published in April 2005 by Peset & Ortín (2005), eighteen in October 2006, and nineteen one month later. With the methodology applied in this study, fourteen more have been identified (almost 50% of the total number). Another necessary aspect is the use of quality filters for the inscription of projects[4]. The recording of only projects already underway, or with a minimum number of data, should be considered, or to purge those projects that have been abandoned after a while. Lastly, the date should refer to the creation of a repository or service, and not to the date of inscription in the record. A considerable number of data and service providers already exist in Spain, Portugal and Latin America. However their contents are disseminated through platforms of a very different nature, as is the case of journals or large digital libraries. At the same time, they consider as projects those that have only a few hundred pieces of data or in some cases no data at all. These aspects do not help in the creation of quality resources, a priority base for the correct promotion and diffusion of Open Access.

The aim of this study is to recommend some good practices not only to improve the current state of OA in the mentioned countries, but also to consolidate and promote the use of new technologies and the access to the whole text in the digital libraries. On the one hand it is necessary to increase the distribution of new contents in whole text and to provide services with increased value. The tools for the federative searches already exist and cover almost all of the records for Spain. Now it is necessary to advance in services as, for example, does ROAI, allowing the visualizing of results in different styles: labelled, ISBD, Author & Title, Dublin Core RDF, Onix, Ibermarc & XML-MARC Schema. It is necessary to touch on the distribution of contents in Hispano-Luso languages. To do so, in conjunction with the development of data providers, it would be desirable to act in the following ways, in harmony with the international evolution:

- Promoting self-archiving. It is necessary to make the authors aware of the importance of diffusing their production. At this time, it is fundamentally the libraries that are leading the development. The authors themselves should be those who take the responsibility for making their contents visible. To do so, the institutions should be helped to establish the mandatory policies which oblige the authors to archive their scientific production which is the result, for example, of public subsidies.

- Supporting with determination the starting and the continuation of the quality projects housed within the institutions. In any case it would not be surprising that, in the short term, the limiting of the number of subsidized data providers should be considered. The aim is to guarantee their maintenance over time, to preserve the data, the quality of the contents, to avoid overlapping and, definitively, to rationalize the cost.
- Promoting coordinated acts at this incipient moment. The diffusion of this initiative and the structuring of the work areas are vital to maximizing the profitability of the investment in this sector. Sharing efforts and knowledge at this time would be a profitable factor for the studied zone. In this way, the development of national, regional or thematic platforms would give a greater visibility to the Luso-Hispanic e-science.

This study aims to be the first step towards the creation of a Reference Web for the OAI initiatives in the Hispano-Luso tongue. OA-LATINO must contain at least the following sections referring to the Open Access sector in Spain, Portugal and Latin America: *Official Descriptive Record OA (refOA), Annual Barometer OA (bOA), Recommendations (recOA)* -Guides to good practices; Tutorials; Checklist to set in motion and to evaluate an archive or a service; Use of common vocabularies, classifications and metadata- and Recommendations for the *use of standards in the authors' signatures* to increase the international impact (iraLIS).

As has been mentioned throughout the text, this meticulous methodology for identification must be spread to Portugal and Latin America. In this way, the overlapping in the distribution of contents could be known in the future, as could the numbers that are offering whole texts, which is the sole motive for the appearance of the Open Access movement.

References

[1] Ayuso, M.D. & Martínez, V. (2006). Evaluación de calidad de fuentes y recursos digitales: guía de buenas prácticas. *Anales de documentación*, 9, 17-42. http://www.um.es/fccd/anales/ad09/ad0902.pdf

[2] Barrueco, J.M. & Subirats, I. (2003). Open archives en las universidades españolas. In *III Workshop de REBIUN sobre proyectos digitales, Barcelona (Spain).* http://eprints.rclis.org/archive/00000384/

[3] Barrueco, J.M. & Subirats, I. (2003). Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH): descripción,

funciones y aplicación de un protocolo. *El Profesional de la Información* 12, 2, 99-106. http://eprints.rclis.org/archive/00000177/

[4] Cobacho, L. (2006). Recursos electrónicos editados por las Bibliotecas Universitarias Españolas: Informe Línea 3 Rebiun. In *I Simposio Internacional "*Bibliotecas y Objetos Digitales*" Madrid 23 a 25 de octubre de 2006*.

[5] Fernández, I. (2005). *El impacto de los archivos de e-prints en la comunicación científica entre investigadores españoles: CSIC aceptación y uso*. Madrid: FESABID.

[6] Ferrer, A., Peset, F., Moreno, M. & Lloret, N. (2005). *Guía metodológica para la implantación de una biblioteca digital universitaria.* Gijon:Trea.

[7] Frías, J. A. & Santiago, T. (2006). Las universidades españolas y la iniciativa OAI: Proyectos consolidados, proyectos en prueba y buenas intenciones. In *Repositorios institucionales: una vía hacia el acceso, la visibilidad y la preservación de la producción científica, Zaragoza (Spain)*.

[8] Harnad, S. & Brody, T. (2004). Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine* 10, 6 http://www.dlib.org/dlib/june04/harnad/06harnad.html

[9] Harnad, S. (2006). Maximizing research impact through institutional and national Open-Access Self-Archiving Mandates. Invited Keynote. *CRIS2006. Open Access Institutional Repositories. Current Research Information Systems. Bergen, Norway, 11-13 May 2006* (In Press). http://ct.eurocris.org/CRIS2006/

[10] Oliver, K.B. & Swain, R. (2006). Directories of Institutional Repositories: Research Results & Recommendations. In *72nd IFLA General Conference and Council, 20-24 August 2006, Seoul* http://www.ifla.org/IV/ifla72/index.htm

[11] Peset, F. & Prats, J. (2006). Interoperatividad entre repositorios y proveedores de servicios. In *Repositorios institucionales: una vía hacia el acceso, la visibilidad y la preservación de la producción científica, Zaragoza (Spain).* http://eprints.rclis.org/archive/00008280/

[12] Peset, F. & Ortín, A. (2005) Tendencias internacionales en el acceso libre a la documentación científica digital : OAI-Open Archives Initiative. (In Press). http://eprints.rclis.org/archive/00004131/

[13] Peset, F., Subirats, I., Barrueco, J. M. & Noverges, N. (2003). Políticas editoriales en la documentación española e implantación

de E-LIS (E-Prints in Library and Information Science). In *Proceedings Contenidos y Aspectos Legales en la Sociedad de la Información (CALSI), Valencia (Spain)*. http://eprints.rclis.org/archive/00000416/

[14] Serrano, J. & Prats, J. (2005). Repertorios abiertos: el libre acceso a contenidos. *Revista de Universidad y Sociedad del Conocimiento (RUSC)*, 2, 2. http://eprints.rclis.org/6599/

[15] Subirats, I. (2004). Introducció a l'accés obert a la literatura científica i tècnica. In *Jornades Catalanes d'Informació i Documentació, Barcelona (Spain)*. http://eprints.rclis.org/archive/00002666/